DOCUMENT RESUME

ED 054 214 TM 000 789

AUTHOR	Kirk, David B.
TITLE	Technique for Approximating the Bivariate Normal Correlation Coefficient, Rho, and Estimating Tetrachoric r.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-35
PUB DATE	Jun 71
NOTE	18p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Algorithms, *Computer Programs, *Correlation, *Mathematical Applications, Mathematics, Probability, *Probability Theory, *Statistical Analysis, Techniques

ABSTRACT

In this paper a reliable method is found for approximating the value of the Bivariate Normal Correlation Coefficient, rho, given values of the joint probability and the normal deviates, h and k, or the related areas. This technique finds useful application in the computation of the tetrachoric correlation coefficient, r, when the underlying distributions may be assumed to be normal. (Author)

RB-71-35

TECHNIQUE FOR APPROXIMATING THE BIVARIATE NORMAL CORRELATION COEFFICIENT, rho, AND ESTIMATING TETRACHORIC r

David B. Kirk

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
June 1971

A Technique for Approximating the Bivariate Normal Correlation Coefficient, rho $\cancel{\phi}$, and Estimating Tetrachoric r

D. B. Kirk

Abstract

In this paper a reliable method is found for approximating the value of the Bivariate Normal Correlation Coefficient, ρ , given values of the joint probability and the normal deviates, h and k , or the related areas. This technique finds useful application in the computation of the tetrachoric correlation coefficient, r , when the underlying distributions may be assumed to be normal.

A Technique for Approximating the Bivariate Normal Correlation Coefficient, ρ, and Estimating Tetrachoric r

D. B. Kirk

In many psychological studies data may be measured in or reduced to a two-variable dichotomy. For example, in a testing situation each item may be scored as correct or incorrect, students may be passed or failed, etc. In order to estimate the correlation between these dichotomies, the assumption is made that the underlying traits are continuous and normally distributed or that they were measured in such a way that a normal distribution could be used as a legitimate model. The data may appear in a form similar to the following 2x2 table:

Variable 2

		Wrong	Right	Totals	Percentage
Variable 1	Right	a	b	a + b	p_1
	Wrong	c	d	c + d	q_1
	Totals	a + c	b + d	n	
	Percentage	q_2	p_2		1

Figure 1

The calculation of the bivariate normal r, or tetrachoric r for the dichotomized case, involves performing an inverse interpolation of the bivariate normal distribution function:

$$(1) \qquad L(h,k,r) = \int_h^\infty \int_k^\infty \frac{1}{2\pi\sqrt{1 - r^2}} e^{-\frac{(x^2+y^2-2rxy)}{2(1-r^2)}} dx\, dy$$

since we are effectively given values of L, the standard deviates h and k, and are required to find r.

In order to note the correspondence of the 2x2 table with the integral, we might consider the cell (Wrong, Wrong), in Figure 1, with a frequency of c or a joint percentage of c/n which corresponds to the value of $L(h,k,r)$. The h and k values are the deviates determined by the areas established by the marginal percentages q_1 and q_2 of Variables 1 and 2 as illustrated below:

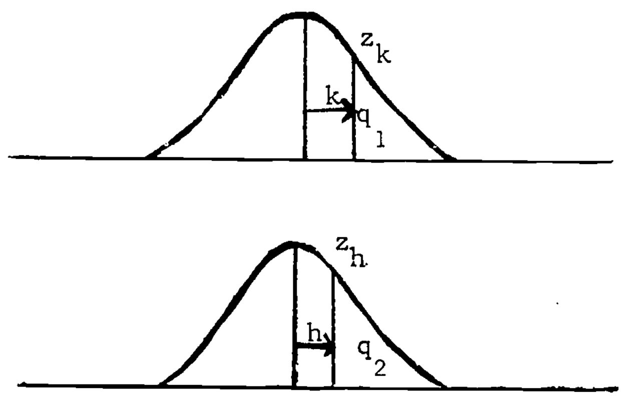

Figure 2

For purposes of consistency throughout the paper, the joint percentage, c/n will be labeled P .

Theoretical approaches to the calculation of r have generally relied on an infinite series approach. A derivation is given in Kendall and Stewart [6] that for the 2x2 table

$$(2) \qquad \frac{\frac{d}{n}}{z_h z_k} \cong \sum_{j=0}^{\infty} \frac{r^j}{j!} H_{j-1}(h) \cdot H_{j-1}(k)$$

where the H_j are the Tchebycheff-Hermite polynomials and the z_h and

z_k are the ordinates on the normal curve as shown in Figure 2. Since, in a dichotomized situation, the percentages are complementary, the use of d/n rather than c/n will introduce a corresponding change in the area calculation.

McNemar's [8] notation includes the restriction that the marginal areas involved are less than or equal to 1/2 but uses the same expansion. His formula is

$$\text{(3)} \qquad \frac{\frac{c}{n} - q_1 q_2}{z_x z_y} = r + xy\,\frac{r^2}{2!} + (x^2 - 1)(y^2 - 1)\,\frac{r^3}{3!} + \ldots$$

This notation will be followed except that we will use h and k instead of x and y to indicate the deviates.

The Hermite polynomials, products of which are used in the expansion as coefficients of $r^n/n!$, are:

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= x \\ H_2(x) &= x^2 - 1 \\ H_3(x) &= x^3 - 3x \\ &\vdots \end{aligned}$$

and the recursion relationship is

$$\text{(4)} \qquad H_n(x) = xH_{n-1}(x) - (n - 1)H_{n-2}(x) \quad \text{for } n \geq 2 \ .$$

Although Kendall warned that series (3) converges very slowly, McNemar indicated this approach would yield reasonable approximations except at the extreme values. Since this procedure was not mentioned in a paper given at the Psychometric Meeting held here at ETS in 1969 on the methods of calculation of the tetrachoric r , it was the first technique programmed in the present study. It should perhaps also be mentioned that IBM's Scientific Subroutine package for the 360 uses the same expansion but limits the series to seven terms. A Newton-Raphson iteration method was programmed and, since many of the calculations of terms in the series could be used both for the function and its derivative, the approach seemed quite sound. Without going into substantial detail about the overflow and underflow problems involved, an attempt was made to calculate r for $h = k = 0$ with a P value of .477473 for which the true value of r is .99. The program finally converged to .995 but required 47 terms in the series. More terms would probably have given increased accuracy, but limits of 10^{+60} and 10^{-60} were specified by the program to prevent the numbers from becoming out of range for the computer. Furthermore, 18 iterations were required to calculate r within a range of .0001 with these 47 terms so, with this rather discouraging information, it seemed desirable to investigate other techniques.

It must be mentioned that for a calculation of this type, with the desirability of examining the output by varying the number of terms, convergence criteria, upper limits of calculation to prevent overflow, etc. and with relatively minimal input and output, the use of interactive computing procedures was virtually a necessity.

At the previously mentioned meeting, a paper on methods of calculation of the tetrachoric correlation coefficient was presented by Ernest C. Froemel [3]. Three methods of calculation were examined and a program written by David R. Saunders using an algorithm by Ledyard Tucker seemed to give the best results computationally but required the greatest computing time (naturally!).

The bivariate normal is rewritten in the form

$$(5) \qquad L(h,k,r) = \frac{1}{2\pi} \int_0^r \frac{1}{\sqrt{1 - x^2}} e^{-\frac{(h^2-2hkx+k^2)}{2(1-x^2)}} dx + q_1q_2$$

where $L(h,k,r) = P$ as defined by our notation. The integral is then approximated by the sum:

$$(6) \qquad \frac{1}{2\pi} \sum_{i=0}^{n} f(x_i)\Delta x \quad \text{where} \quad f(x) = \frac{1}{\sqrt{1 - x^2}} e^{-\frac{(h^2-2hkx+k^2)}{2(1-x^2)}} .$$

The value of Δx is fixed at .0078125 and successive summations are made until the sum equals $P - q1q2$ as determined by a change of sign. The value of $n \cdot \Delta x$ then approximates the value of r. The approximation of the integral by a linear trapezoidal technique is rather fundamental. However, it is direct, simple, easy to understand, and avoids problems of discontinuity and overflow and underflow. As a possible improvement, one might be inclined to use Simpson's rule as a curvilinear approximation and hope for equivalent accuracy with fewer intervals. However, for our purposes, Saunders' existing program was converted to double precision and yielded the following results for $h = k = 0$:

Table 1

Saunders' Program

P	Computed r	True r	No. terms required.
.315495	.399999	.40	52
.411699	.84999	.85	109
.428217	.899988	.90	116
.477473	.989857	.99	127

By virtue of the Summation Method, the discontinuity problem evident in other techniques is avoided and a numerical result is assured. This may be at the expense of additional computing time for a given level of accuracy, however. Since the range from 0 to 1 is divided into 128 partitions ($\Delta x = \frac{1}{128}$), the number of terms should never exceed 128.

However, since $f(x)$ is a smooth function over the range involved one would hope that adequate accuracy for the integral might be achieved by a shorter method with associated savings.

Since it is necessary to adjust and converge on the unknown upper limit, assuming we are within an interval of convergence, the Newton-Raphson method provides a rapidly converging technique. Gaussian quadrature, since it provides good accuracy with relatively few unequally spaced points, will be used to evaluate the integral.

Thus letting

$$f(r) = \frac{1}{2\pi} \int_0^r \frac{1}{\sqrt{1 - x^2}} e^{-\frac{(h^2-2hkx+k^2)}{2(1-x^2)}} dx \tag{7}$$

we need $f'(r)$. It is shown in Courant [2] that if

$$F(x) = \int_{g_1(x)}^{g_2(x)} f(x,y)\, dy$$

then

$$F'(x) = \int_{g_1(x)}^{g_2(x)} \frac{\partial}{\partial x} f(x,y)\, dy - g_1'(x) f(x,g_1(x)) + g_2'(x) f(x,g_2(x)) \quad .$$

Consequently,

$$f'(r) = \frac{1}{2\pi\sqrt{1 - r^2}}\, e^{-\frac{(h^2-2hkr+k^2)}{2(1-r^2)}} \tag{8}$$

which should give little computational difficulty except for $|r|$ close to 1.

To employ Gaussian quadrature, and restrict the upper limit of the integral to 1, a scaling or variable transformation $u = x/r$ is made. Then $x = ur$, $dx = rdu$ and the integral becomes:

$$f(r) = \frac{r}{2\pi} \int_0^1 \frac{1}{\sqrt{1 - u^2r^2}}\, e^{-\frac{(h^2-2hkur+k^2)}{2(1-u^2r^2)}}\, du \tag{9}$$

$$\cong \frac{r}{2\pi} \sum_{i=0}^{n} w_i g(u_i) \quad \text{where} \quad g(u) = \frac{1}{\sqrt{1 - u^2r^2}}\, e^{-\frac{(h^2-2hkur+k^2)}{2(1-u^2r^2)}} \tag{10}$$

in which the u_i are the roots of the Legendre Polynomials, and the w_i are the associated weights for an $(n + 1)$ point quadrature.

After a starting value is determined, successive values are computed by the Newton-Raphson iteration method:

$$r_{i+1} = r_i - \frac{f(r_i) - m}{f'(r_i)} \tag{11}$$

where $m = (P - q1q2)$. Iteration is continued until

$$|r_{i+1} - r_i| < \varepsilon \quad .$$

In the current program version, a five-point quadrature is used with the following values for u_i and w_i ,

Table 2

Roots of Legendre Polynomials and Associated Weights

i	u_i	w_i
0	.04691008	.1184634
1	.23076534	.2393143
2	.5	.2844444
3	.76923466	.2393143
4	.95308992	.1184634

The convergence criteria, ε , on the successive r_i is .0001 and the number of iterations is limited to 50.

Various techniques for establishing a starting value of r were investigated, since experimentation showed that not only the speed of convergence but actual convergence itself was dependent upon a reasonable starting estimate, even though the derivative is less than 1. The value finally used:

$$r_{est} = \frac{P - q_1 q_2}{z_h z_k}$$

was taken from the first term of the series expansion, and was restricted to lie between the limits of -.97 and .97. This approximation works satisfactorily for most of the cases. However, if the first attempt fails, an arbitrary r value of .55 is used as a starting value for a final computation.

Using this technique, the test cases for $h = k = 0$ yielded the following results:

Table 3

Gaussian Quadrature-Newton Raphson for $h = k = 0$

$\underline{P}$	Computed r	True r	Iterations
.315495	.40003	.40	2
.411699	.85006	.85	4
.428217	.90015	.90	4
.25	0	.00	1
.477473	.9949	.99	6

It is evident that by using only a 5 point quadrature and from 1 to 6 iterations we have reasonable results (to 3 decimals except when $r > .99$), and we are performing substantially fewer calculations than the Saunders' program.

Consequently, with reasonable success at the $h = k = 0$ level (for which $L(0,0,r) = 1/4 + \arcsin r/2\pi$ exists as a closed solution), Hastings' approximation [4, p. 192] was coded to calculate h and k from the given areas. The following table illustrates the accuracy of that subroutine:

Table 4

Hastings' Approximation for h and related z_h

Area (q) (Input)	h (calculated)	h (true)	z_h (calculated)	z_h (true)
.5	-1.01×10^{-7}	0	.39894228	.39894228
.158655254	.999968	1	.24197835	.24197072
.022750132	2.000435	2	.0539440	.0539910
.001349898	3.000314	3	.00442768	.00443185

Finally, using the above routine to calculate h and k and the quadrature-iteration technique for r , we have the following examples for positive and negative values of r near the extreme values where one would naturally expect the most trouble.

Table 5

Hastings Gaussian Newton-Raphson Iteration Results

P	h	k	r (calculated)	r (true)	Iterations
.25	0	0	0	.00	1
.079328	0	1	3.86×10^{-6}	.00	1
.011375	0	2	-3.07×10^{-6}	.00	1
.000675	0	3	2.89×10^{-5}	.00	1
.00061	3	3	.9003	.90	4
.000031	2	3	.0012	.00	1
.158631	0	1	(see below)[a]	.95	
.022742	1	2	(see below)[b]	.95	
.001349	2	3	(see below)[c]	.95	
.000809	3	3	.9510	.95	4
.477473	0	0	.9949	.99	6
.2420389	0	0	-.0500	-.05	1
.0505413	0	0	-.9507	-.95	4
.0007048	0	1	-.9005	-.90	5

[a,b,c](Using 8 point Gaussian quadrature and slightly more accurate estimates for h and k , these values converged to .94961, .9502, and .9511 respectively.)

Note that difficulty occurs when P is extremely close to the area under the normal curve (as shown in Table 4) for either the h or k . This corresponds to nearly equivalent cell and marginal percentages in the 2x2 diagram which further implies one of the cells has nearly zero frequency. Difficulty will also occur for P values extremely close to zero.

Conclusion

This study has shown that Gaussian Quadrature supplemented by a Newton-Raphson iteration technique provides a rapid method by which a reasonable estimate of tetrachoric r may be obtained. Difficulty may occur when marginal percentages and P values are extremely close or for P close to zero.

Since the satisfactory performance of any entity is dependent upon satisfactory performance of the components comprising that entity, it seems worthwhile to examine the major components of this program.

1. The Gaussian Quadrature.

   Only 5 points were used in this study which is really a rather coarse mesh. 10 points will certainly give better accuracy, and 40 point quadrature is not uncommon. Naturally, this will be at the expense of computing time and at some point may become less efficient than the Saunders.' technique. Additional experiments comparing accuracy vs. time may be made at a later time.

2. Estimates of h and k from the Hastings' approximation.

   These values also may be made more accurate by an iteration technique. This probably should be done for critical computations.

   For example, a P of .022742 ($h = 1, k = 2$, $r = .95$) did not converge with the current version of this routine, but converged to .9501 in 8 iterations using exact values for h and k. For most practical applications, however, it is hoped that three decimals will suffice for h and k.

3. The Convergence Criterion.

   ε was set at .0001 for the examples cited in this paper. Although this may be tightened, it is necessary to realize that the process is merely converging upon the estimate of r as computed by the number of points specified in the Gaussian quadrature and not the true r. It is obviously inadvisable to use an extremely small convergence criterion with coarse quadrature.

4. The starting value of r.

   From the study, it is known that convergence to a solution is contingent upon a reasonable starting estimate. However, this sensitivity is probably due more to truncation problems than estimates falling outside an interval of convergence.

The Program

A listing of the program provides the additional, necessary, unambiguous documentation required to complete the paper. It is, after all, thi program, supporting the analysis, which provides the numerical results.

To reduce compilation costs and increase speed, it was written in BASIC and programmed on IBM's CALL360 system. The complete program, except for exponential, logarithmic, and square root routines provided by the system, is listed at the end of the paper. A translation to FORTRAN is a simple task for a reasonably experienced programmer.

Input. The input typed in at the console consists of the P value, the marginal percentages q_1 and q_2 (both $\leq .5$), and a test parameter (1 or 0) to indicate whether iterative calculations are to be or not to be printed.

Output. If convergence is achieved, the result "OK," tetrachoric r, h and k, and the number of iterations required are printed.

If r becomes greater than 1, a flag is set, and the calculation is repeated with a different starting estimate. A second failure causes a return to the read statement. At this point the same data may be re-entered with the test parameter set to 1 to investigate the cause of the failure.

The Computer Program

```
100 INPUT A1,B1,B2,T
110 X=0
120 F = B1
130 GOSUB 790
140 H = E3
150 Z1=E4
160 F = B2
170 GOSUB 790
180 K=E3
190 Z2=E4
200 A = (A1 - B1*B2)
210 A3 = A*6.283185307
220 R2 = A/(Z1*Z2)
230 IF R2 > .97 THEN 260
240 IF R2 < -.97 THEN 280
250 GO TO 290
260 R2 = .97
270 GO TO 290
280 R2 = -.97
290 IF T=0 THEN 310
300 PRINT"A,A1,A3,B1,B2,H,Z1,K,Z2,R2";A,A1,A3,B1,B2,H,Z1,K,Z2,R2
310 FOR I=1 TO 50
320 U=1
330 GOSUB 700
340 P1=M6
350 U=.04691008
360 GOSUB 700
370 P2 = .1184634 * M6
380 U = .23076534
390 GOSUB 700
400 P3 = .2393143 * M6
410 U = .5
420 GOSUB 700
430 P4 = .2844444*M6
440 U = .76923466
450 GOSUB 700
460 P6 = .2393143 * M6
470 U = .95308992
480 GOSUB 700
490 P7 = .1184634 * M6
500 P5 = R2 * (P2 + P3 + P4 + P6 + P7)
510 R3 = R2 -(P5-A3)/P1
520 R5=R2
530 R4 = ABS (R2 - R3)
```

The Computer Program (continued)

```
540 IF T = 0 THEN 560
550 PRINT "R2,R3,P1,P5";R2,R3,P1,P5
560 IF R4<.0001 THEN610
570 R2 = R3
580 NEXT I
590 PRINT "FAILED TO CONVERGE" R5,R3
600 GO TO 100
610 PRINT "OK";R2,H,K,I
620 GO TO 100
630 X = X+1
640 PRINT"M1 NEG,U,R2,R3,P5,P1,A,H,K,M1,I"
650 PRINT U,R5,R3,P5,P1,A,H,K,M1,I
660 IF X=2 THEN 100
670 PRINT "LAST TRY, R = .55"
680 R2 = .55
690 GO TO 290
700 M = U*R2
710 M1 = 1-M*M
720 IF M1<0 THEN 630
730 M2=2*M1
740 M4 = -(H*H + K*K - 2 * H * K * M)
750 M5 = SQR (1/M1)
760 M8=EXP(M4/M2)
770 M6 = M5 * M8
780 RETURN
790 IF F>.5 THEN 860
800 E = SQR(-2.*LOG(F))
810 E1 = ((.010328*E) + .802853) * E + 2.515517
820 E2 = (((.001308*E)+.189269)*E + 1.432788)*E+1
830 E3 = E - E1/E2
840 E4 = .39894228*EXP(-E3*E3/2)
850 GO TO 880
860 PRINT "F>.5";F,B1,B2
870 GO TO 100
880 RETURN
890 END
```

Bibliography

[1] Chesire, L., Saffir, M., & Thurstone, L. Computing diagrams for the tetrachoric correlation coefficient. Chicago: University of Chicago Press, 1933.

[2] Courant, R. Differential and integral calculus. New York: Interscience Publishers, 1956.

[3] Froemel, E. Paper presented at Psychometric Society Spring Meeting, Princeton, N. J., April, 1969.

[4] Hastings, C. Approximations for digital computers. Princeton, N. J.: Princeton University Press, 1955.

[5] IBM System 360 Scientific Subroutine Package (360A-CM-03X); CALL/360 Time Sharing System. White Plains, N. Y.: International Business Machines, 1970.

[6] Kendall, M., & Stewart, A. The advanced theory of statistics. New York: Hefner, 1961.

[7] Kunz, K. Numerical analysis. New York: McGraw-Hill, 1957.

[8] McNemar, Q. Psychological statistics. New York: Wiley, 1955.

[9] Mood, A. Introduction to the theory of statistics. New York: McGraw-Hill, 1950.

[10] National Bureau of Standards. Tables of the bivariate normal distribution function (AMS-50). Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1959.

[11] National Bureau of Standards. Handbook of mathematical functions (AMS-55). Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1964.